THE MEASUREMENT OF INTELLIGENCE BY THE BINET TESTS.

By Cyril Burt.

Part II.

(*Continued from* Vol. VI., No. 1, p. 36.)

Hitherto we have been concerned with the object of the Binet scale. We have seen that Binet aimed at determining the normal course of the development of intelligence and at measuring individual differences in developmental terms. We have examined his conception of intelligence and his conception of mental development. We have inferred that, from the very nature of these conceptions, his scheme could only fulfil his aims in a very tentative and limited way.

We may now turn to consider the actual achievements of the scale; and examine with what success these aims have actually met. The results obtained by different investigators will, I think, when examined side by side, confirm our previous inferences.

(1) We will examine, first of all, the success with which the scale of tests has been arranged. It will be remembered that one of the striking features of the Binet system is the adoption of what I have called the principle of external gradation. Each test is not chosen so as to provide, of its own intrinsic nature, a graded scale of units inherent in itself. Rather, each test is itself a unit; and each marks a single point in an ascending scale. Thus, " the scale is composed of a series of tests, increasing in difficulty, commencing on the one hand at the lowest intellectual level that can be observed, and reaching at the other

that of an average and normal level."[1] In table III. of his last article, Binet has "arranged the tests according to their difficulty, the more easy being placed before the more difficult, and the degree of difficulty being indicated by the figures given. . . . This table (he adds) is to be retained to judge the results obtained by other observers ; it is a norm."[2] A study of the table shows that Binet's own results do not quite correspond with the order given. According to the figures, the test of "definition superior to use" is passed by only 23 children out of 40; yet it is placed before three other tests passed by 29 or 30 children out of 40, and indeed by more children at every age except one. Here, therefore, Binet seems to have determined the relative ease of the tests upon a priori grounds. The discrepancy he apparently would regard as merely analogous to "the discrepancy between a theoretical curve and an experimental one." Other observers' orders show wider deviations still. Their figures are seldom given in a form which allow of close comparisons. So far, however, as I can calculate, the correlations between the arrangements suggested by other investigations and that of Binet range from about ·9 to about ·7. The discrepancies shown by some of the tests are considerable. On rearranging the tests in order of difficulty according to results obtained in London schools, the test of the "divided oblong," which is the fourteenth test in Binet's list, appears as twenty-fifth; "colour-naming," which is treated as a much harder test, and is placed twenty-fourth in Binet's list, appears as only sixteenth. There is a yet more serious difficulty. However accurately we make our calibration, it appears that the order of difficulty for normals will never be the same as that for defectives. Thus, normal children apparently come to define objects in terms of use a year before they can count thirteen pennies; with defectives the reverse appears to be the case. Defectives are disproportionately backward in comparing objects, arranging weights, describing pictures and counting in reverse order. With epileptics the arrangement has still further to be revised. In

[1] *L'Annéé Psychologique,* 1905, p. 194.

[2] *L'Année Psychologique,* 1911, p. 153.

the tests involving weight-discrimination, memory, motor response, and rapidity of thought epileptics are unusually handicapped.

What then are we to do when the fixed points upon our scale vary from observer to observer, and become interchanged as we pass from normals to defectives and epileptics?

(2) But, perhaps, in demanding that the order of different observers' arrangements should be identical as regards each single test, we are demanding too much. We should merely require that the several groups of tests should be invariably assigned to the same mental age. Let us, therefore, see how successful Binet has been in discovering the age at which each test should be passed.

Binet suggested two criteria for verifying the correct allotment of the tests to their respective ages, neither being very rigid. His first requirement is that the number of children shown by the tests to be backward shall be equal to the number shown to be forward. The results of Binet and others only conform to this requirement if we mass the results obtained from children of all ages together; but even at its best the conformity is not strict. In Binet's results, out of 264 children, 72 are below the age and only 63 above. In Goddard's results, out of 1,547 children, 599 are below age and only 394 above. So far both results suggest that, if anything, the tests are a little too hard. Let us turn from the classical investigations of France and America to less known researches in Italy and Russia. Of 144 normal children from a kindergarten and State schools of a thickly populated part of Rome, only 17% are below age and over half, namely 54%, are above age. Yet, of 187 normal workmen's children attending State schools in Moscow, none are above age, and 73% are below; of the most capable children 3% were found to be two years behindhand, and, therefore, according to Binet would fall under an extremely strong suspicion of feeble-mindedness. Tested by the criterion proposed, the original standardisation of such a scale plainly has only a very local value; quite apart from the impossibility of finding equivalent translations for the questions and equivalent values for the coins, the simple transference of a scheme devised for children of Parisian labouring classes to examine children of English elementary schools would be actually misleading unless the scheme is first re-standardised for differences of race and again re-standardised for differences of sex and social class.[1]

Meanwhile, we must note that the criterion proposed for testing the standardization itself calls for preliminary study. As used by Binet and his followers it is open to at least two objections. First, the criterion clearly assumes a "normal," or, at any rate, a symmetrical distribution of ability. This is an assumption which most urgently calls for investigation; at present, apart from the general convenience of the assumption, the grounds for its acceptance are extremely scanty. At least one psychologist of note, Professor Cattell, has maintained the contrary. After pointing out that, in schools and colleges, selection would tend to yield a curve of distribution skewed in the negative direction, he writes: "In spite of this factor, I believe that the main skew is in the opposite (positive) direction; and that ability is distributed something like wages which are roughly proportional to it." In crude language, dullards outnumber geniuses, just as paupers outnumber millionaires. If this belief be true, then the results of Binet and Goddard are sound; but the criterion which seems to indicate their unsoundness is itself unsound.

There is a second objection to the common mode of applying this criterion. Piling together all the ages obscures the very fact we want to know. How accurately are the tests assigned to *each* age? Calculate the number above and below age separately; and then an approximation to symmetry is the exception rather than the rule. Terman and Childs have

[1] A curious instance of racial or social differences is afforded by the list of contradictory statements prescribed for the detection of absurdities. Binet's series (relating to a trivial railway accident with 48 killed, the mutilated corpse of a supposed suicide, the choice of a luckier day than Friday for killing oneself) were found rather blood-curdling by Whipple, and accordingly replaced. Binet protests. "Il parait que cest histoires semblent effrayantes aux jeunes Americains. Nos jeunes Parisiens en ont ri."

According to the most recent investigations (J. and R. Weintrob, "The Influence of Environment upon Mental Ability as Shown by the Binet and Simon tests," *Journal of Educational Psychology*, 1912, p. 576), the effect of difference in social class appears to be small. In any case, it may well be due to differences in hereditary mental ability. I do not, therefore, press this oft-exploited argument against the scale.

done this. They find that the tests are far too easy for the earlier ages, and far too hard for the later. Out of 83 five-year-old children, only one is below age; 77 are above. Out of 35 twelve-year-old children, 32 are below age and none are above. Similar results have been repeatedly observed. Clearly, by adding the figures for all ages together such inequalities hide and compensate for one another.

Binet, however, hints at another criterion. Namely, a test is too easy for a given age, if nearly all can perform it; it is too hard if nearly half fail. His statements, however, of his grounds for allotting each test to the ages chosen are extremely vague:[1] once more we are left with the impression that the scheme was first drawn up in the study on the basis of clever guesswork and rough trials, and corrected and re-corrected in the light of subsequent experience. Later investigators have realized the need for a more exact formulation of the criterion. Some investigators require each test to be passed by two-thirds of the children of the age to which it is to be assigned[2]; others by rather over a half[3]; others again, by nearly all.[4] Seventy-five per cent. is perhaps the only figure that has more than one supporter. Assuming normal distribution, it has been suggested that half of a given group (that is, all within the limits of the probable error) might be supposed to have medium or "average" ability; and that, of the remaining half, one-half again would be above "average," and one-half below. All those of "average" or above "average" ability should be required to pass the test, that is, in all, three-quarters of the entire group. At present, these suggestions, however plausible, have very little basis in experimentally ascertained fact.

Let us for the moment accept this arbitrary criterion. Let us compare the ages indicated for each test in the half dozen

[1] In the articles his phrases are: "all the children" . . . or "almost all the children could do this test"; or again, "this test was passed by the majority of the children of this age." In a letter to Dr. Bobertag, he states that "a test may be assigned to a given age if only 65 per cent. succeeded; . . . . if 90 per cent. succeed it is perhaps too easy."

[2] Terman and Childs. These writers, however, propose to modify their criticisms in cases where there is a sudden rise in successes obtained in passing from one year to the next or where there is a similar proportion of successes in several successive years.

[3] Pearson and Jaederholm calculate that in particular cases the solutions by children of the right age, according to a rearranged scale, were only 55 per cent.

[4] The requirement of Goddard and his followers is 75 per cent. *or more.*

most important investigations with the scale. In every case the age chosen is to be that at which 75% of a group of normal children pass the test. The discrepancies are amazing. For not one test is there complete agreement as to the age to which it should be assigned. Among the forty tests, the commonest result, occurring in some fourteen cases, is for a test to be assigned now to the same age as Binet's, now to the next above, and now to the next below—thus fluctuating over a range of three years. In only seven cases is the range limited to two years. Twelve tests vary over a range of four years; six over a range of five years; two over a range of six years.[1] And yet a child who varies by two years or more from the normal age as given by the tests is, according to the scheme, under a grave suspicion of mental deficiency!

(3) Let us, however, assume that by dint of extensive investigations upon normal children, the several tests have been accurately assigned to their respective ages; and that the scale has, therefore, been properly standardized and fixed. How successfully will it enable us to measure the degree of intelligence possessed by any given child?

Three methods have been proposed. First, we may calculate his "absolute mental age." To do this, it would seem sufficient to carry him up the scale as far as he can go, and then give the age corresponding to the last group of tests which he passes with complete or almost complete success. This plan, however, Binet eventually modified. A child does not break down at one definite point; his failures may be spread over a series of mental years. Hence, Binet's final suggestion was to take the last age at which the child passes all the tests successfully, and then add on a fifth of a year for every further test passed beyond that point. Other writers have suggested that we should also deduct a fifth of a year for failures committed

[1] The table published by Meumann, *loc. cit.,* Vol. II., p. 276, gives one test (problem-questions) as extending over a range of seven years, being assigned to the age of 15 by Binet, and to the age of 9 by Goddard. The 9, however, seems to be an error. The statements in the text are based on figures compiled not by myself, but by an American reviewer (J. C. Bell, *Journal of Educational Psychology,* Vol. III., p. 104-5). His statement that "there is a surprising agreement in the results of the different investigators" is sufficient to acquit him of bias against the Binet scale.

before that point; others, that we should weight the tests to be added according to their difficulty.[1]

As Binet remarks, his final mode of calculation permits us to measure the mental age even to fractions of a year: but the fractions "do not deserve an absolute confidence." For the classification of mental defects he suggests the following plan. Debiles or feeble-minded children are those who can read and write, but cannot achieve complex or abstract thought; they, therefore, are said to correspond to a mental age of five to nine. Imbeciles possess the use of speech, but cannot communicate by means of reading or writing; they, therefore, are said to correspond to the mental age of two to five. Idiots do not possess the use of language; and they, therefore, have a mental age below that of the normal child of two.

We cannot, however, class a child of eight as feeble-minded because he cannot perform the tests assigned to the age of nine. Binet, therefore, suggested a second form of measurement, that of "mental retardation." The child's mental age is subtracted from his physical age, and the remainder measures his degree of backwardness. A child who is three years behind the normal standard of his age is considered mentally defective; a child who is two years behindhand falls under extremely strong suspicion of feeble-mindedness; a retardation of but one year has, in this respect, little or no significance. To these proposals there are several objections. First, a child's chronological age is by no means the simple unambiguous measure it at first sight appears. By "aged 7" one investigator means from 7 to 7$\frac{11}{12}$ years inclusive; another means between 6½ and 7½ years; another means from 6 to 6½ years; yet another

[1] The importance of these suggestions will become obvious on considering the following scores, each of which yields the same mental age according to the usual method of reckoning. (1 indicates success; 0, failure.)

Child.	Tests for				
	Age VII.	Age VIII.	Age IX.	Age X.	Age XI.
A. B.	1 0 0 0 0	1 1 1 1 1	1 1 1 0 0	0 0 0 0 0	0 0 0 0 0
C. D.	1 1 1 1 1	1 1 1 1 1	1 1 1 0 0	0 0 0 0 0	0 0 0 0 0
E. F.	1 1 1 1 1	1 1 1 1 1	1 0 0 0 0	0 0 0 0 1	0 0 0 0 1

Scores inclining to types A. B. and E. F. are especially common among epileptics; C.D. is always rare.

from $6\frac{10}{12}$ to $7\frac{2}{12}$ years inclusive. It is not at all uncommon, especially in the case of backward and defective children to find that the age borne by them throughout school life apparently alters when the time for leaving draws near or when a birth certificate has been produced. In many ways it would be far preferable to use some verifiable measure of physical age, more immediately related to physical growth, such as height, or degree of ossification of cartilage (Rotch's improved X-ray method); or (were it practicable) degree of pubescent or pubertal change. Disparity between mental age and physiological age is far more significant than disparity between mental age and chronological. Secondly, a large proportion of children diagnosed upon other grounds as mentally deficient prove to have only slight degrees of retardation. Out of 236 children in a German special school, 88 were either "normal" or backward by only one year according to the 1908 scale. None, however, have a mental age of over 9. Similar results have recently been obtained in English special schools. Finally, Binet has not after all been successful in finding a measure which is independent of age. A retardation of two years is commonly described as though it meant the same thing at any age of life. But clearly a retardation of six years at 30 would not have the gravity it possessed at 13; at 3 it would, from its nature, be unobtainable. In all special schools that have been investigated the older children generally show a larger retardation: in an English school I find that there is a correlation of ·43 between retardation and chronological age. In considering retardation, therefore, we have to take into account the chronological age just as in considering the absolute mental age. Proper allowance could only be made by the method of correlation: by calculating, for instance, from experimental data obtained with a large group of defectives the regression of mental retardation upon their chronological age, and then correcting in the usual way. So far as I am aware this has never been done.

A third measure has, however, been proposed by Professor Stern, the "intellectual quotient." Here the mental age is divided by the physical age, instead of being subtracted from it.

We thus have an estimate of the proportion of the amount of intelligence a child has to the amount of intelligence he ought to have. If this falls below four-fifths, *i.e.*, 0·8 (corresponding to a retardation of one year at the age of 5 and of three years at the age of 15), then, it is supposed, we have evidence of mental deficiency. But even this measure has been found to be dependent in part upon absolute age; indeed, after the age of 20 even a normal individual's intellectual quotient must get rapidly smaller. And the few investigators who have tried this mode of measurement have not met with great success.[1]

From the nature of these proposals it must be clear that a further topic calls urgently for investigation. It is not sufficient to know the general course of mental development of normal children : we need also to know the course of mental development among defectives. Does it consist of a temporary arrest followed by an advance at the normal rate? Or of a normal advance followed by a premature arrest? Does it consist in a rate of advance slower than normal throughout? Or in a rapider onset of the gradual decline in rate? Or, finally, are different courses followed by different forms of defect? Until these questions are investigated, it is impossible to decide upon general grounds between intellectual quotient, degree of retardation, absolute mental age, or any other method proposed for measuring the degree of a defective's development in relation to normal child's development at the same age.

In conclusion, let us take these methods of estimating of individual intelligence simply as arbitrary measures, and consider their empirical value upon merely practical grounds. Without enquiring into their validity let us compare them directly with the only other measure we have—namely, the subjective estimate of a careful and conscientious daily observer. In short, let us calculate, for a given group of children, the correlation between the teachers' ranking for intelligence and the order of intelligence yielded by the tests. Curiously enough, this has hardly ever been done. In a school of about a hundred defectives, I find that the correlation of the teachers' estimates with absolute mental age is ·55; with mental retarda-

[1] Jennings and Hallock, "The Binet-Simon Tests at the George Junior Republic," *Journal of Educational Psychology*, 1913, p. 8.

tion, ·59; with intellectual quotient, ·48. Almost the only other calculations of this relationship, with which I am acquainted, are those obtained from normal children in Scotland by Dr. McIntyre and Miss Rogers. In the abstract of a paper read before the psychological subsection of the British Association (Birmingham, 1913), they state that " the indices of correlation range from ·85 to ·16, the majority being about ·5."[1] Now, internally graded tests which give no higher correlation with imputed intelligence than ·5 are, as a rule, at once rejected, as no tests of intelligence at all. A simple test of some " higher mental process," taking from two to five minutes, for instance the " opposites," " analogies," or Ebbinghaus' " completion " tests, usually gives after two applications only, correlations of about ·7 or ·8. A combination of several such internally graded tests will give a more satisfactory order still.[2]

Tested, therefore, both in theory and in practice, the Binet scale proves far less satisfactory than is commonly claimed. The limited and tentative character of his schemes Binet repeatedly emphasized. Many of the difficulties urged he himself foresaw. It is not his work that I criticise. It is rather the current tendency to take his work, against his express injunctions, as a final and finished product that I deprecate. As a provisional but practicable plan for testing mental deficiency, as a rough but intelligible method of interpreting the results, as a pioneer investigation of the general course of mental development, as a demonstration of the richness of the higher, more complex, and more ordinary mental processes, as a protest against the mere examination of acuity of sensation, of speed of reaction, or of anatomical peculiarities, as a means of interesting the teacher, the doctor, and the social worker in the measurement of psycho-

[1] Mr. Dumville has recently obtained from a small normal group a correlation of a similar order, viz., " According to Spearman's foot-rule, ·43, or translated into Pearson's coefficient, ·62." (*Journal of Experimental Pedagogy*, Vol. 2, No. 2.)

[2] cf., e.g., Vickers and Wyatt, " Grading by Mental Tests," *Journal of Experimental Pedagogy*, Dec., 1913. Unfortunately, no satisfactory tests of higher mental processes, applicable to very young children or defectives have as yet been published. I have, however, obtained fairly promising results with complex "substitution" and "erasure" tests, after first modifying them so as to use more interesting material—figures of little men and animals instead of the usual letters and geometrical orms.

logical capacities by psychological devices, as a prolific source of inspiration and suggestion, and, finally, as a stimulus to scientific discussion and enquiry, in these and many other ways the Binet scheme remains a marvel and a masterpiece. But every work of genius calls for later readjustments before it can be exploited as a practical instrument. Binet himself was always the first to modify his plan in the light of other investigators' research.

The most recent work has, I fancy, indicated modifications, perhaps more drastic than even he anticipated. Two surrenders will, I think, have to be made.

First, for all exact and scientific purposes, the principle of external gradation, of constructing a scale out of a long list of heterogeneous tests arranged in order of their relative difficulty, will have to be given up. Within the limits contemplated it seems impossible to find an order of difficulty which shall be the same for all. Further, the plan of "one test one point" throws open the door to chance. All Binet's tests were alternative tests: the child either succeeds or fails. Consequently, either he or the examiner is faced with a dilemma—the one situation in a psychological experiment which most invites the play of chance. The child is asked: "Which is your right hand?" "Is it morning or afternoon ?" A correct response may be due to a blind choice of the first alternative that occurred almost as often as to genuine knowledge. Or, again, the child copies a square or a diamond; and the examiner has to decide whether it is a fair reproduction or not. Such decisions are bound to be arbitrary and unreliable. Where a time limit of 6, 10, or more seconds is allotted, and the child is failed if he takes longer than the prescribed allowance, the results in borderline cases are apt to be more irregular and haphazard still. Each test, therefore, must be made to provide its own scale. The measure must be, not simply failure or success, but so many problems correctly solved within the given time, or so many seconds taken to complete the task. The actual number might well be registered mechanically by the use of more ingenious apparatus; and the whole performance be rendered more independent of the power of examiner and examinee to understand instructions.

Secondly, we must discard the principle of measuring intelligence in terms of age. Each mental capacity should be measured in units of its own. These may be, first, the natural and original units of the test, expressed in seconds, marks, or other convenient form. For a rough illustration of their significance we may relate the measure thus obtained with the nearest age-norm. But it cannot be expected that it will coincide precisely with the average for a given year, or much less for fifth of a year. Better still, we may convert the original measurements into terms of the variability of the group. The mean for the corresponding age may then be taken as zero, and the probable error or standard deviation may be used as unit. We can then estimate at once the probable frequency of any given measurement or the likelihood of its occurring in a normal population of a given size; and the units in various parts of the scale will be far more nearly equivalent. Finally, by the aid of yet further calculations, based upon correlation, we may devise an index which shall measure general intelligence independently of age or of the nature of the tests employed.

The simplicity of these calculations is perhaps worth illustrating. Suppose we desire to determine the most probable measure of the intelligence of a child whose performance at a given test is 14. We may assume the following constants to be known: $r = \cdot 6$ represents the correlation between the test and intelligence; $\bar{x} = 20$ represents the average performance of the child's group at the given test; with $\sigma_1 = 4$ as standard deviation; $\bar{y} = 30$ represents the average measure of the intelligence of the group; with $\sigma_2 = 7$ as standard deviation. Then, the measure required, the measure of the child's intelligence is given by the usual formula, $y - \bar{y} = r \frac{\sigma_2}{\sigma_1} (x - \bar{x})$.

Substituting the known values, $y - 30 = \cdot 6 \times \frac{7}{4} (14—20)$; and, therefore, $y = 23\cdot 7$. This value, however, is only true within certain limits; but even these limits can be determined. The value is really the average of an array of possible values, whose standard deviation $S_y = \sigma_y \sqrt{1—r^2} = 7 \sqrt{1—(\cdot 6)^2} = 5\cdot 6$.

So far we are upon recognised ground.[1] Following Binet's repeated injunctions, however, we shall use not one test, but several; let us say, for simplicity, three. Here I would propose to apply regression equations calculated by means of the formulae for multiple correlation. The equations will be of the form

$$x_1 = 0\cdot 127\, x_2 + 0\cdot 587\, x_3 + 0\cdot 034\, x_4$$

where x represents the most probable value of the intelligence of a child, whose performances or marks at three tests are x_2, x_3, x_4. The constants by which the marks are multiplied are determined from the partial correlations between intelligence and the tests. Thus, $0\cdot 127 = b_{12\cdot 34} = r_{12\cdot 34} \frac{\sigma_{1\cdot 234}}{\sigma_{2\cdot 134}} = 0\cdot 68 \frac{9\cdot 17}{49\cdot 2}$ where, $r_{12\cdot 34}$ is the partial co-relation between intelligence (1) and a given test (2) with the other tests (3, 4) constant. This in turn can be determined from the original total correlations between intelligence and first test ($r_{12} = \cdot 49$), first test and second test ($r_{23} = \cdot 15$) and so on.

[1] I have taken this instance with some modification from W. Brown, *Essentials of Mental Measurement*, p. 46.

Using three simple tests, "Finding Opposites" (O), "Completing Syllogisms" (S), and "Completing Argument" (A), and estimating both Intelligence (I) and test-performance in terms of ranks for convenience, I have obtained the following equation from a group of 60 normal children:

$$I = \cdot 60\,(A) + \cdot 23\,(O) + \cdot 17\,(S).$$

Judged by teachers' estimates of intelligence, this furnishes a far better measure of intelligence than either the best test taken alone, or the average of all three unweighted. When amalgamated by this procedure three or four of the best Binet tests give far better results than ten or fifteen when amalgamated on the principle of each test to count the same. But the determination of such equations calls for much fuither research.

This, or something analogous, is, in my judgment, the only way to obtain a single measure of general intelligence from a variety of tests. We should constantly apply such statistical methods to the returns of the Census or of the Board of Trade; yet in the case of mental ability or defect we are content with raw and uncorrected estimates. In any case, correlation is essential to indicate kinds of tests most closely related to intelligence, to select the forms of those tests which are most reliable and self-consistent, and, finally, to solve the more fundamental and prior problems as to the nature of general intelligence and of the various specific mental capacities.

Last of all, it has become increasingly clear that we need not one, but several scales, each carried not merely to the age of thirteen, but extended through puberty and adolescence to the cessation of mental growth. The tests of scholastic attainment, the tests of general knowledge, the tests of emotional and moral character must not, as in the Binet schemes, be mixed with the tests of intelligence and other simpler psychical capacities. Further, the tests of the several specific capacities must be kept distinct from each other. Each has its own development; and each must have its own independent scale. Binet himself has drawn up a scale for testing general knowledge. Thorndike, Ayres, and Courtis, in America, have drawn up scales for measuring scholastic abilities—handwriting, arithmetic, and literary composition. In the work of Freud and Jung we have the beginnings of a scheme of emotional and moral tests. It remains for English investigators to complete the list. Along these lines only can we hope to do justice to the incredible variety both of mental ability and of mental defect.